

TESTING THE MEASUREMENT EQUIVALENCE OF INTERACTIVE VOICE RESPONSE (IVR) AND PAPER VERSIONS OF THE EQ-5D

J. Jason Lundy, PhD and Stephen Joel Coons, PhD
Center for Health Outcomes and PharmacoEconomic Research
The University of Arizona, Tucson, AZ 85721-0202, USA

REFERENCES

Agel J, Greist JH, Rockwood T, et al. Comparison of interactive voice response and written self-administered patient surveys for clinical research. *Orthopedics* 2001;24:1155-1157.

Alemi F, Stephens R, Parran T, et al. Automated monitoring of outcomes: application to the treatment of drug abuse. *Medical Decision Making* 1994;14:180-187.

Coons SJ, Gwaltney CJ, Hays RD, et al. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Research Practices Task Force Report. *Value in Health*. In press.

Crow JT. Receptive vocabulary acquisition for reading comprehension. *Modern Languages Journal* 1986;70:242-50.

Henriksen B. Three dimensions of vocabulary development. *Studies in Second Language Acquisition* 1999;21:303-17.

Macran, S. Test-retest performance of EQ-5D. In Brooks R, Rabin R, de Charro, F (eds). *The Measurement and Valuation of Health Status using EQ-5D: A European Perspective*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2003:43-54.

McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1996 ;1(1):1, 30-46 (Correction, Vol. 1, No. 4, 390).

Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* 1979;86:420-8.

Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use* (3rd ed.). New York: Oxford University Press, 2003.

US Food and Drug Administration. *Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims (DRAFT)*, February 2006. Available at: <http://www.fda.gov/cder/guidance/5460dft.pdf> [Accessed June 1, 2008].

ABSTRACT

OBJECTIVE: Electronic data capture technologies, such as interactive voice response (IVR) systems, are emerging as important alternatives for collecting patient-reported outcomes data. The objective of this study was to assess the measurement equivalence of an IVR version of the EQ-5D with the original paper version.

METHODS: This study utilized a crossover design with subjects randomly assigned to one of two assessment orders: 1) paper then IVR or 2) IVR then paper. A convenience sample of in-treatment outpatient cancer clinic patients (n=139) were asked to complete each assessment two days apart. The analyses tested for mean differences (repeated measures ANOVA) and reliability (intraclass correlation coefficient [ICC]) to assess measurement stability over time. Equivalence of the means was established if the 95% confidence interval (CI) of the mean difference was within the minimally important difference (MID) interval: -.035 to .035 for the index and -3 to 3 for the EQ VAS. Adequacy of the ICC was established by comparing the ICC 95% lower CI with a critical value of 0.70.

RESULTS: The per protocol analysis included 109 subjects for the EQ VAS and 113 subjects for the index. For the EQ-5D index, the means (SD) of the paper and IVR administrations were 0.790 (0.172) and 0.800 (0.180), respectively. The 95% CI of the mean difference was -0.024 to 0.006, which was within the equivalence interval. The ICC was 0.894 (95% lower CI 0.857), significantly different from 0.70. For the EQ VAS, the means (SD) were 72.0 (19.7) for paper and 74.1 (19.8) for IVR. The 95% CI of the mean difference was -3.784 to -0.484, partially within the equivalence interval. The ICC was 0.897 (95% lower CI 0.859) also significantly different from 0.70.

DISCUSSION: This analysis provides evidence that the EQ-5D scores on the IVR version were equivalent to those obtained on the original paper version.

INTRODUCTION

BACKGROUND: Most health status and health-related quality of life measures were originally developed to be administered via paper-and-pencil. However, advances in technology have provided more efficient methods for data collection and database creation. Although there are clear advantages to collecting data electronically, equivalence between the alternative modes of data collection must be established. For use in clinical trials, the US Food and Drug Administration (FDA) is now requiring such evidence when a patient-reported outcome (PRO) measure is modified or adapted to a new mode of administration (e.g., electronic device) (FDA 2006).

Modifications that occur when an instrument is adapted from paper to an electronic platform include changes in the wording or placement of instructions, wording or order of the items and/or their response options, length of a visual analog response scale, as well as adaptation from visual cognitive processes to aural cognitive processes. This latter modification, specific to

telephone-based administration modes, may constitute one of the more significant departures from the original paper-based mode of administration.

Telephone-based voice/aural devices are commonly referred to as interactive voice response (IVR) systems. IVR systems interact with callers using a pre-recorded voice question and response system. Some of the advantages of IVR systems are that no additional hardware is required for the respondent other than a telephone, little if any respondent training is necessary, and data are stored directly to the central database. In addition, the use of the recorded voice prompts has been shown to reduce the literacy skill requirements of study participants (Crow 1986; Henriksen 1999).

OBJECTIVE: There is a lack of peer-reviewed literature addressing the measurement equivalence of IVR-based versions with the original paper-based, self-administered versions of PRO questionnaires (e.g., Alemi et al. 1994; Agel et al. 2001). Hence, a study comparing the measurement equivalence of IVR administration of a widely used instrument, namely the EQ-5D, to the original paper versions is warranted and timely. The purpose of this research was to assess the measurement equivalence between the original paper-based version and an IVR version of the EQ-5D.

METHODS

SAMPLE SIZE CONSIDERATIONS: This study tested for mean differences as well as differences in intraclass correlation coefficients (ICCs); however, the approaches for determining the sample size differ. We opted for the more conservative sample size calculation (i.e., 110 subjects) which is provided from the computation of the sample requirements for our tests of the ICC (Streiner and Norman 2003).

INCLUSION CRITERIA: All subjects were cancer survivors 18 years of age or older and currently in treatment (for cure or palliation). Treatment included chemotherapy, radiation, a combination of both, or other medical treatments. Further, the subjects must have had access to and the ability to use a touchtone phone, as well as an understanding of written and spoken English.

SUBJECT RECRUITMENT: Subjects were recruited through the Arizona Cancer Center's outpatient clinics in Tucson. Interested individuals had the opportunity to enroll based upon face-to-face contact with a study recruiter at the clinics. Also, individuals who learned about the study from our flyers were able to enroll by calling a dedicated phone line. The study was conducted under the auspices of the University of Arizona's Human Subjects Protection Program. All subjects who agreed to complete the study questionnaires received a \$20 gift card.

RANDOM CROSSOVER DESIGN: The use of the crossover design in this study involved the random assignment of respondents to complete either a paper questionnaire or the IVR-based questionnaire for the first administration and then the other mode for the second administration. Testing



College of Pharmacy



ClinPhone is now part of the Perceptive Informatics® family.

Acknowledgements: The data used for this research were collected as part of a study funded by ClinPhone Plc. Additional staff and facility support necessary to conduct this research was provided by the University of Arizona, College of Pharmacy and the Arizona Cancer Center's Behavioral Measurements Shared Service. The views expressed in this paper are those of the authors and do not necessarily represent the views of ClinPhone Plc or the University of Arizona.

and order effects are threats to the validity of this design, but the within-patient design (i.e., subjects as their own control) provides greater statistical power and decreases sample size requirements. Figure 1 describes how the crossover design was employed in this study.

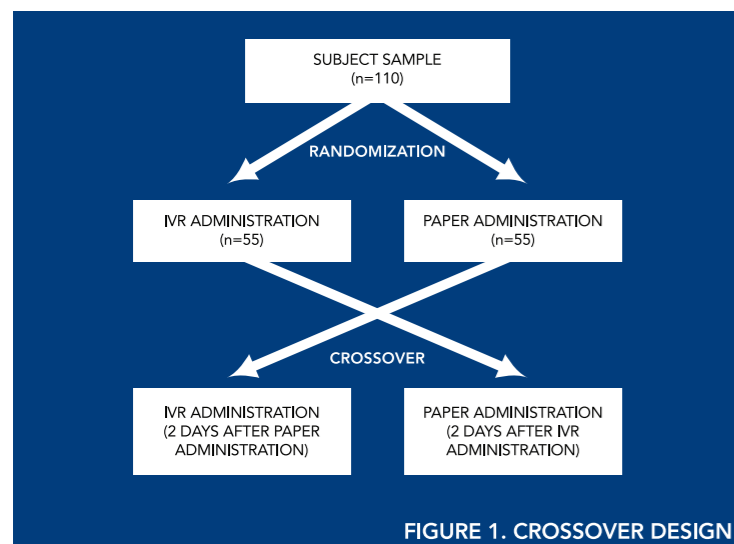


FIGURE 1. CROSSOVER DESIGN

After random assignment, a study packet was mailed to each participant. The packet contained a cover letter, a study information sheet, the study disclosure form and two sealed envelopes labeled "1" and "2." Each of these two envelopes was labeled with the date on which they were to be opened for self-administration of the questionnaire. The two administrations were scheduled two days apart during a relatively stable phase of their course of treatment. The subjects were contacted on the day of the first scheduled administration to remind or confirm completion of the first questionnaire and encourage them to complete the second when appropriate.

Each envelope contained information regarding completion of the designated mode of administration. The envelope instructing participants to complete IVR administration included written information regarding accessing the IVR system using their phone. The envelope instructing participants to complete the self-administered paper version of the questionnaire contained a self-addressed postage-paid return envelope. The cover letter asked the subject to complete the paper questionnaire and mail it back upon completion.

If there was a delay in receipt (via mail) or completion (via IVR) of the questionnaires, the recruiters called to inquire about any difficulties the subject may have encountered and/or to confirm that the subject was fully aware of the study protocol timelines and encourage completion of the tasks.

STUDY MEASURE: The EQ-5D self-report questionnaire consists of two parts: a descriptive system and a visual analog scale. The descriptive system comprises the following five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. For each dimension, respondents are asked to describe their own health with respect to three levels, reflecting "no problems," "some problems," and "extreme problems." These responses are used to classify the respondent into one of 243 unique EQ-5D health states. A scoring function assigns a value to self-reported health states from a set of preference weights that have been empirically derived. The resulting EQ-5D index score is on a scale where 1.0 represents perfect health and 0.0 represents death.

In addition to the descriptive system, the EQ-5D also has a visual analog scale (EQ VAS) to measure an individual's overall health status. The original EQ VAS is a thermometer-like 20-cm vertical line with endpoints labeled "worst imaginable health state" and "best imaginable health state" anchored at 0 and 100, respectively.

The EQ-5D descriptive system was adapted to the IVR system using the exact wording for the items and responses. For the EQ VAS, the IVR system asked respondents to "picture in their minds" a scale with 100 at the top (i.e., "best health state you can imagine") and 0 at the bottom (i.e., "worst health state you can imagine") and enter a number between 0 and 100 representing their health status.

DATA ANALYSIS: Descriptive statistics of cancer type, age, and sex were calculated to characterize the sample; no subgroup analyses were performed. All statistical analysis were performed using SPSS version 16.0 (Chicago, Illinois) and evaluated using an alpha level of 0.05.

Mean Differences: Testing of the mean differences was based on analysis of variance (i.e., split-plot ANOVA) with factors for mode, period of administration (first or second) and subject; the p-values from the significance tests will be reported. The split-plot ANOVA also accounts for the interaction effect (period x mode effect, often termed as carryover). The adjusted mean differences between modes were estimated together with the associated 95% confidence interval (CI) for the difference. Equivalence on this measure was considered to have been established if the 95% CI excludes the MID used in the sample size calculations, namely 0.07 for the EQ-5D index and 6 points for the EQ VAS.

Reliability: To analyze the reliability of the instrument, the analyses were based upon the ICC (Shrout and Fleiss 1979). It approaches 1.0 when the between-groups effect is very large relative to the within-groups effect. ICC is 0 when within-groups variance equals between-groups variance, indicative of the grouping variable having no effect. The ICC is calculated based on the ANOVA model that includes factors for mode and subject. A one sided 95% CI for the lower bound was computed using the formula provided in McGraw and Wong (1996). Measurement equivalence was considered to have been established if the lower bound of the 95% CI exceeded 0.70.

RESULTS

SUBJECTS: A total of 184 subjects agreed to participate. Of those, 139 subjects completed both administrations for a response rate of 75.5% (Figure 2). The respondents were 67.6% female and had a mean age of 61.5 years. The ages ranged from 19 to 86.

The analyses for the index score and the EQ VAS were based on a "per protocol" analysis. Hence, subjects were in the analyses if they completed both questionnaire administrations within 72 hours. Furthermore, subjects were excluded from the analysis if their score difference exceeded two standard deviations, namely 40 points on the EQ VAS and 0.28 on the EQ-5D index score. The analysis included 109 subjects for the EQ VAS and 113 subjects for the EQ-5D index.

DATA QUALITY: The amount of missing and unusable data from the paper and IVR administrations was tabulated for all subjects who completed both questionnaire administrations. Of the 139 paper responses, only two missing responses, from the same subject, were noted on the descriptive system. Similarly, only two missing responses were observed, from different subjects,

on the IVR version of the descriptive system. There were four subjects out of the 139 (2.9%) who did not complete the paper EQ VAS in contrast to no missing data on the IVR version of the EQ VAS. Further, there were 53 subjects (38.8%) who did not complete the paper EQ VAS according to the instructions. Twenty-four subjects (17.3%) drew a line from the bottom of the VAS scale (i.e., originating from '0') to the point representing the valuation of their health. Thirteen subjects (9.4%) drew a circle around the corresponding VAS response and an additional 13 subjects drew a line across the VAS to select their value. Three other unusual responses were noted: two subjects drew an arrow pointing to their VAS response and one subject placed an 'X' on the scale.

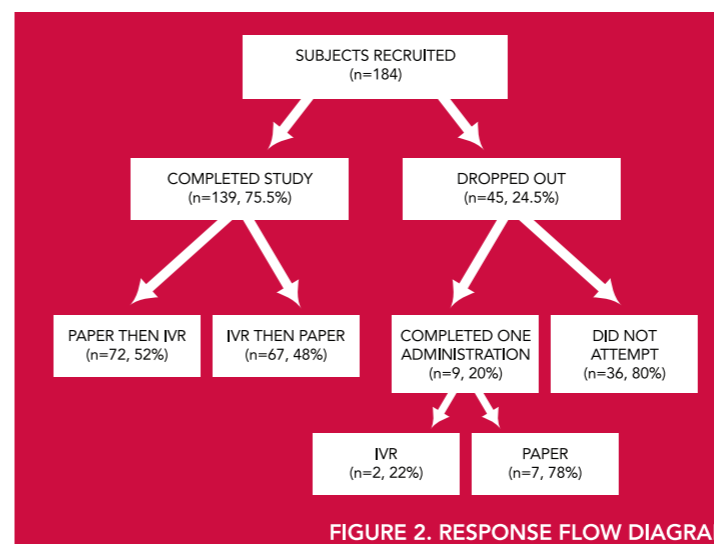


FIGURE 2. RESPONSE FLOW DIAGRAM

While no missing responses were noted on the IVR version of the EQ VAS, this mode was not free of unusual phenomenon. There were eight subjects who entered a single digit response on the IVR EQ VAS. Four of those single digit responses corresponded exactly with the 1st digit of the two digit VAS response given on paper. For example, a subject may report a EQ VAS score of 70 on the paper version but the IVR version score was recorded as a 7. Two subjects reported paper VAS scores of 70 and 40 but the corresponding IVR scores were 8 and 5, respectively. Of the remaining 2 subjects, one subject did not respond to the paper VAS and the other subject had very disparate scores of a 40 on the paper and a 7 on the IVR version.

MEAN DIFFERENCES: For the EQ-5D index, the means (SD) of the paper and IVR administrations were 0.790 (0.172) and 0.800 (0.180), respectively. The tests for an order effect and for an order by mode interaction based on the split-plot ANOVA were not statistically significant. The adjusted means (i.e., least squares means [SE]) were 0.789 (0.016) for the paper version and 0.798 (0.017) for the IVR version. The adjusted mean difference was 0.009 and the 95% CI of the mean difference was -0.024 to 0.006, which was within the equivalence interval.

The EQ VAS means (SD) were 72.0 (19.7) for paper and 74.1 (19.8) for IVR. Similarly, no order effect or mode by order interaction was present in the analysis of means for the EQ VAS. The adjusted means were 72.03 (1.85) for the paper and 74.16 (1.89) for the IVR. The adjusted mean difference (SE) was 2.13 (0.83) and associated 95% CI of the mean difference was -3.784 to -0.484, partially contained within the equivalence interval of -3 to +3.

DIMENSION	% AGREEMENT	KAPPA (SE)
MOBILITY	81.3%	0.657 (0.074)
SELF-CARE	95.0%	0.761 (0.115)
USUAL ACTIVITIES	79.1%	0.650 (0.065)
PAIN/DISCOMFORT	84.9%	0.721 (0.064)
ANXIETY/DEPRESSION	86.3%	0.746 (0.059)

TABLE 1. RESPONSE AGREEMENT FOR THE EQ-5D DIMENSIONS

RELIABILITY: The ICC was 0.894 (95% lower CI 0.857), significantly different from 0.70. Furthermore, the percent of exact agreement and kappa coefficients for each of the five dimensions of the descriptive system are provided in Table 1.

DISCUSSION

These results provide substantial evidence supporting the measurement equivalence of the paper and IVR versions of the EQ-5D. The mean difference CI for the EQ-5D index score reflected equivalence of the means from the two modes, as the CI was wholly contained inside the equivalence interval.

In contrast, the mean difference CI for the EQ VAS was only partially contained in the equivalence interval, providing an inconclusive result of neither equivalence nor non-equivalence of the mean VAS scores. However, both the index score and the EQ VAS had ICCs or kappa coefficients that were on par or higher than those observed in the literature for the test-retest comparison of the paper version (Macran 2003). As stated in the recommendations from ISPOR's ePRO Task Force (Coons et al. In press), electronic modes of administration should not be held to a higher standard than the original paper-based version.

These results are limited by the lack of generalizability to the general population or the overall cancer patient population. Since internal validity of equivalence studies is of high importance, we feel this was an acceptable tradeoff.

Further, we encountered a small amount of unusable data from the IVR version of the EQ VAS. We observed the phenomenon of a subject entering a single digit, such as a 9, upon one administration and having a value such as 90 on the other administration. It is unlikely that the subjects intended such a wide disparity when evaluating their current health status, particularly over the two- to three-day study interval. This situation results in unusable data provided by the IVR system. To overcome this limitation, we recommend that when items with response sets exceeding single digits (e.g., EQ VAS) are adapted to IVR systems, the system should prompt the subject to confirm his or her choice. This solution should be easily incorporated.

CONCLUSION

The evidence presented here, when taken in totality, supports the measurement equivalence of the IVR version of the EQ-5D with the original paper version. In conclusion, we recommend that the data from this IVR version of the EQ-5D be treated as equivalent to that of the paper version.